



Jurnal Ilmiah Informatika (JIF)

| ISSN (Print) 2337-8379 | ISSN (Online) 2615-1049 |

Jurnal online di akses di <http://ejournal.upbatam.ac.id/index.php/jif>



SCRAPING WEB MARKETPLACE MENGGUNAKAN METODE DOM PARSING UNTUK PENGUMPULAN DATA PRODUK

Muhammad Joko Umbaran Haris Bahrudin^a Hardan Gutama^b

^a Universitas Alma Ata, Sistem Informasi, Indonesia

^b Universitas Alma Ata, Informatika, Indonesia

INFORMASI ARTIKEL

Sejarah Artikel

Diterima Redaksi: 28 Februari 2020

Revisi Akhir: 10 Maret 2020

Diterbitkan Online: 25 Maret 2020

KATA KUNCI

Web, Web Scraping, DOM, Parsing, Marketplace

KORESPONDENSI

E-mail: haris.bahrudin@uaa.ac.id

A B S T R A C T

Web Scraper is a way of extracting scripts that run there that are commonly chosen to do web memos via DOM parsing. Specific nodes that are collected using DOM parsers and tools like XPath help the process of scraping web pages. In this study using the DOM Parsing method to obtain product data on the market. Scraper can be further developed as a data collection technique on the internet for further research that can be provided about the concept of big data to be used for forecasting or getting the information needed. Alone Parsing is a way of breaking data or symbols, both in language and in language, according to formal grammar rules. After using the parsing technique, it will be generated again using Parse Tree, which is a process of compiling product data that forms like a tree, using syntak analysis to break down product categories.

1. PENDAHULUAN

Web scraping saat ini menjadi tren yang banyak dilakukan oleh perusahaan atau pun individu untuk tujuan tertentu. Teknik merupakan metode untuk mengekstraksi data dari halaman website. Anda bisa saja secara manual mengcopy detail data dari halaman web ke halaman spreadsheet, namun biasanya data yang ada di dalam website merupakan data yang besar sehingga membutuhkan tempat berkapasitas besar serta waktu yang cukup lama. Oleh karena itu, salah satu cara yang dapat Anda gunakan untuk mengunduh data besar dari website adalah dengan menggunakan "web scraper". Web scraper adalah program yang dapat membuka halaman website kemudian mendownload data yang ada di dalam web, mengekstrak ke dalam format yang terstruktur, dan menyimpannya ke dalam sebuah file atau database. Web scraper dapat mengunduh konten yang biasanya berupa teks dan diformat sebagai HTML dari beberapa halaman web dan mengekstrak data darinya. Dalam penelitian ini scraping produk menjadi hal yang populer yang digunakan untuk menjual produk yang sama, kebanyakan para pelaku dropshiper menggunakan teknik scraping untuk menjual produk yang sama di toko online. Pada penelitian ini Scraping menggunakan metode DOM Parsing secara umum digunakan mengetahui cara kerja internal halaman web dan mengekstrak skrip yang berjalan di dalamnya biasa memilih untuk melakukan

web scraping melalui parsing DOM. Node spesifik dikumpulkan menggunakan parser DOM dan alat-alat seperti XPath membantu proses scraping sebuah halaman web.

2. TINJAUAN PUSTAKA

Terdapat 3 Metode yang biasanya digunakan untuk scraping website salahsatunya adalah HTML Parsing merupakan metode yang paling sering digunakan dalam proses parsing data dari halaman website. Pada umumnya, HTML parsing dilakukan menggunakan JavaScript dan menargetkan halaman HTML linear dan nested. Script ini digunakan untuk mengekstraksi tulisan, link dan data. Dan yang kedua adalah Regular Expressions metode ini berguna jika Anda ingin melakukan tugas ekstraksi data yang sederhana. Sebagai contoh seperti ketika Anda ingin mendapatkan daftar semua email dari halaman web. Regular Expressions ini tidak cocok untuk pekerjaan ekstraksi yang rumit, seperti mengekstrak data dari beberapa halaman deskripsi produk di situs web E-commerce. Namun akan sangat berguna untuk proses transformasi dan pembersihan data. Dan yang terakhir adalah DOM parsing yang digunakan dalam penelitian ini merupakan metode yang paling cocok untuk mengambil data pada webiste karena mempunyai beberapa metode pokok yang bisa menjadi penunjang untuk pengambilan data.

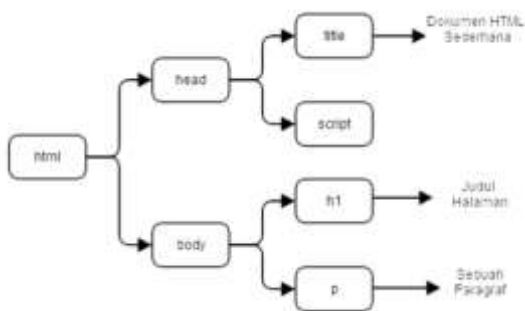
3. METODOLOGI

Pada penelitian ini menggabungkan dua teknik DOM (Document Object Model) dan Parsing. DOM adalah Konten, style, dan struktur file XML, HTML dan SVG yang bersifat cross platform dan merupakan bahasa yang independen. Penggunaan DOM biasanya digunakan untuk javascript dengan fungsi mengubah tampilan web secara dinamis. Parsing Sendiri merupakan Suatu cara memecah data atau serangkaian simbol, baik dalam bahasa alami atau dalam bahasa komputer, sesuai dengan aturan tata bahasa formal. Setelah menggunakan teknik parshing maka akan diturunkan lagi menggunakan Parse Tree yaitu merupakan suatu proses kompilasi data produk yang berbentuk seperti pohon digunakan analisa siyntak untuk memecah kategori produk.

DOM

Document Object Model (DOM) merupakan sebuah ketentuan yang dikembangkan oleh W3C untuk berinteraksi dengan objek-objek yang ada di dalam HTML, XML, maupun XHTML. DOM bersifat cross-platform dan language-independent, yang artinya DOM dapat digunakan dengan bahasa pemrograman yang berdiri sendiri, dalam sistem operasi manapun. Tentunya pengembang bahasa pemrograman atau sistem operasi harus mengimplementasikan antarmuka DOM terlebih dahulu sebelum dapat kita gunakan pada aplikasi kita.

Standar DOM dikembangkan untuk berinteraksi dengan elemen-elemen dokumen HTML dan XML, mulai dari pembuatan elemen baru sampai dengan manipulasi dan penghapusan elemen. Pada bagian ini kita akan membahas konsep-konsep dasar untuk berinteraksi dengan DOM tanpa masuk ke pembahasan masing-masing elemen dalam DOM. Fokus pembahasan akan ditujukan ke pemanfaatan DOM dengan efektif.



Gambar 1. Struktur Tree DOM

Struktur *Tree* seperti pada gambar di atas merupakan cara DOM melihat dokumen HTML. Perhatikan juga bahwa terdapat dua jenis elemen pada pohon: node (simpul) yang ditampilkan dalam bentuk kotak putih, serta teks yang ditampilkan dalam bentuk tulisan. Setiap elemen HTML adalah merupakan node, dan dapat membungkus elemen HTML lainnya. Elemen teks, di sisi lain, tidak dapat memiliki elemen lain di dalamnya.

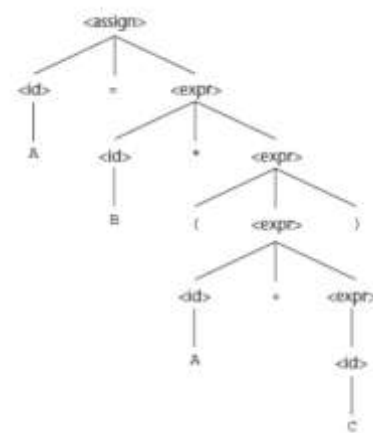
Scrapers yang ingin mengetahui cara kerja internal halaman web dan mengekstrak skrip yang berjalan di dalamnya biasa memilih untuk melakukan web scraping melalui parsing DOM. Node spesifik dikumpulkan menggunakan parser DOM dan alat-alat seperti XPath membantu proses scraping sebuah halaman web.

Parsing

Metode parsing Parsing atau proses penurunan adalah Analisis parsing atau sintaksis adalah proses menganalisis serangkaian simbol, baik dalam bahasa alami atau dalam bahasa komputer, sesuai dengan aturan tata bahasa formal.

Parse Tree

Pada tahap parse tree ada satu metode lagi yang digunakan yaitu menggunakan *Top-Down parsing* adalah langkah dalam membangun sebuah parse tree berdasarkan input dimulai dari root dan membuat nodes untuk parse tree secara priode. Penelusuran dari root ke leaf atau dari simbol awal ke simbol terminal. Pembangunan Parse Tree ini didasarkan pada Grammar yang digunakan. Apabila Grammar yang digunakan berbeda, maka Parse Tree yang dibangun harus tetap berdasarkan pada Grammar yang berlaku.



Gambar 2. Parse Tree

Rumus Parse Tree :

$$A = B * (A + C)$$

Grammar adalah alat generative untuk mendefinisikan bahasa. Sentence dari bahasanya di tentukan melalui urutan aturan aplikasi. Proses parsing menggunakan aturan-aturan yang ada pada Grammar, kemudian membandingkannya dengan kalimat yang diinputkan. Struktur paling sederhana dalam melakukan parsing adalah Parse Tree, yang secara sederhana menyimpan rule dan bagaimana mereka dicocokkan satu sama lain. Setiap node pada Parse Tree berhubungan dengan kata yang dimasukkan atau pada nonterminal pada Grammar yang ada. Setiap level pada Parse Tree berkorespondensi dengan penerapan dari satu rule pada Grammar.

DOM Parsing

Untuk mengetahui cara kerja internal halaman website dan mengekstrak script yang berjalan di dalamnya, Anda dapat melakukan web scraping menggunakan parsing DOM (Document Object Model) . Dengan bantuan web browser, progam dapat mengakses dynamic content dari script client-side yang sudah dibuat.

academic hyperlink creation. Information research, Vol. 8, No. 3, pp. 8-3.

[4]. Bar-Ilan, J. (2015). What do we know about links and linking? A framework for studying links in academic environments. Information Processing & Management, Vol. 41, No. 4, pp. 973-986.

[5] Smith, A. & Thelwall, M. (2017). Web impact factors for Australasian universities. Scientometrics, Vol. 54, No. 3, pp. 363-380.

[6] Noruzi, A. (2019). The Web Impact Factor: a survey of some Iranian university web sites. Studies in Education & Psychology

[11] Smith, A.G. (2019). The Impact of Web sites: a comparison between Australasia and Latin America. Proceedings of INFO, Vol. 99.

[7] Jati, H. (2019). Web Impact Factor: a Webometric Approach for Indonesian Universities. In International Conference for Informatics for Development. Yogyakarta

[8] Rowlands, Ian. (2019) The internet: its impact and evaluation. Proceedings of an international forum held at Cumberland Lodge, Windsor Park, 16-18th . London ASLIB/INTI. Hal 126.

[9] B. Haris .M, “Analisis webometrics ranking universitas negeri dan swasta di indonesia menggunakan web impact factor,” 2017.

BIODATA PENULIS



Penulis Pertama

Muhammad Joko Umbaran Haris Bahrudin
Saya Merupakan Dosen Dan Juga Praktisi
Dibidang Internet Marketing. Karirnya
Dimulai Tahun 2015 Sebagai Analis
Website Di Balaikota Semarang.
Dilanjutkan Dengan Bergabung Di 3G
Production Sebagai Divisi Marketing.
Haris Bahrudin Merupakan Lulusan S2
Universitas Diponegoro Angkatan 2016.



Penulis Kedua

Deden Hardan Gutama atau biasa
dipanggil Hardan adalah seorang pengajar,
manajer website di Universitas Alma Ata,
serta praktisi User Experience Research
dan praktisi Digital Marketing. Hardan
merupakan lulusan S2 dari Universitas
Amikom Yogyakarta angkatan 2016.